Erschienen in: : Brückner, Dominik (2014): Die Qual der Wahl: Google Bücher und die Selbständigkeit des Systems. In: Andrea Bambek und Volker Harm (Hg.): Fremd- und Lehnwortschatz im sprachhistorischen Wörterbuch. Hildesheim [u.a.]: Olms (Germanistische Linguistik - 228), S. 197–213.

DOMINIK BRÜCKNER

Die Qual der Wahl: Google Bücher und die Selbstständigkeit des Systems

1. Einführung

Nach dem Tod von Gerhard Strauß im Jahr 2006 wollte die Redaktion des Deutschen Fremdwörterbuchs (DFWB) am Institut für Deutsche Sprache in Mannheim den dadurch eingetretenen Einschnitt unter anderem dazu nutzen, die Quellenbasis des Wörterbuchs zu erweitern. Bis zu dieser Zeit arbeitete die Redaktion nahezu ausschließlich aus den über 2 Millionen Belegzetteln ihres eigenen Archivs sowie mit den über COSMAS[1] zugänglichen Texten. Onlinequellen waren nur spärlich erschlossen worden.

Kurz darauf startete die Firma Google ihr kostenloses Online-Tool *Booksearch*, heute zu Deutsch *Google Bücher*[2]. Da zu Anfang noch vergleichsweise wenige deutschsprachige Texte in Google zu finden waren, wurden die dort zu Verfügung gestellten Quellen nur selten und ergänzend herangezogen; nach der Bereitstellung von Digitalisaten der Bayerischen Staatsbibliothek (seit 2008) jedoch in immer größerem Umfang.

Die anfängliche Begeisterung machte schnell einer Ernüchterung Platz: Zu den bereits sehr schnell bemerkten Problemen des Systems traten mit wachsender Kenntnis immer weitere, die erst bei längerer, intensiver Nutzung sichtbar wurden. Dies führte bald zu der Erkenntnis, dass die Nutzung der Google-Quellen nur mittels klarer und strenger Richtlinien sinnvoll und zielführend sein kann.

---

[1] Angaben zum Link siehe Literaturverzeichnis unter COSMAS.

[2] Angaben zum Link siehe Literaturverzeichnis unter GOOGLE BÜCHER.

Grundlegende Fragen und Probleme von *Google Bücher* aus lexikographischer Sicht wurden bereits an anderer Stelle diskutiert,[3] so dass im Folgenden einige der weniger deutlich erkennbaren, aber darum umso gravierenderen Probleme von *Google Bücher* in den Fokus gerückt werden können. Dabei soll der Art und Weise, wie *Google Bücher* an verschiedenen Stellen einer Suche eine eigenständige Auswahl trifft, die vom Benutzer nicht nachvollzogen oder gar beeinflusst werden kann, besondere Aufmerksamkeit gewidmet werden.

Die folgenden Beobachtungen ergaben sich aus der täglichen Arbeit am DFWB, und sind notwendigerweise an den Bedürfnissen ausgerichtet, die sich aus dieser Arbeit ergeben. Nichtsdestoweniger ist jedoch zu hoffen, dass diese Erkenntnisse auch für andere lexikographische Arbeiten fruchtbar gemacht werden können.

2. Arbeitsmöglichkeiten

Die Suchoptionen der „erweiterten Buchsuche“[4] dürften hinlänglich bekannt sein, darum soll an dieser Stelle ein kurzer Überblick genügen: Möglich ist eine Suche nach Texten, in denen sämtliche eingegebene Suchbegriffe unabhängig von der Reihenfolge ihrer Eingabe vorkommen („mit allen Wörtern“), nach Texten, in denen alle Suchbegriffe in der Reihenfolge ihrer Eingabe vorkommen („mit der genauen Wortgruppe“) sowie nach bis zu 32 verschiedenen Suchbegriffen („mit irgendeinem der Wörter“). Zudem lassen sich Texte aus dem Suchergebnis ausschließen, in denen ein eingegebener Begriff mindestens einmal vorkommt („ohne die Wörter“). Daneben gibt es optionale Sucheinschränkungen, die es ermöglichen, nach Sprache, Titel, Autor, Verlag, Veröffentlichungsdatum oder ISBN/ISSN zu filtern. Deren Nutzen ist allerdings stark von der OCR-Qualität abhängig und zudem auch aus anderen Gründen stark eingeschränkt: So funktioniert weder der

---

[3] Zu grundlegenden Fragen s. BRÜCKNER (2009 und 2012a).

[4] http://books.google.de/advanced_book_search.

Sprachfilter befriedigend, noch kennt das System den Unterschied zwischen einem Autor und einem Herausgeber.

Für den historisch arbeitenden Lexikographen dürfte die (in krudem Deutsch formulierte) Suchoption „Veröffentlichungsdatum Büchern mit Veröffentlichungsdatum zwischen" besonders interessant sein. Sie ermöglicht die zeitliche Einschränkung einer Suche durch die freie Eingabe zweier Jahreszahlen, und damit unter anderem die Suche nach Früh- und sog. Erstbelegen, das Entdecken neuer (historischer) Sublemmata, das gezielte Ersetzen von Wörterbuch-Buchungen durch Belege aus Primärquellen, das Füllen von Beleglücken oder das Ergänzen eines bislang nur in zu eng geschnittener Form (etwa auf einem Belegzettel) vorliegenden Belegtexts.

3. Was leistet Google Bücher in dieser Hinsicht?

Zur Möglichkeit der Suche nach Früh- und „Erstbelegen" wurde bereits in Brückner (2009 und 2012a) einiges gesagt, etwa anhand der Beispiele *Hobby*, das etwa 150 Jahre vor dem bislang vermuteten Entlehnungszeitraum nachgewiesen werden konnte[5], *Gouvernementalität* und *Homöopathie*, für die ähnliches gilt.[6] Auch zur Unterscheidung zwischen (vielleicht von Wörterbuch zu Wörterbuch weitergetragenen) Okkasionalismen und tatsächlich gebrauchten Wörtern ist dort anhand von Beispielen wie *Gagist*[7], *helikopterisch*[8], *Holokauste*[9], *Hum-*

[5] Wie einige der Frühbelege noch in der Form *Hobby-Horse*, s. DFWB 7, 328.

[6] DFWB 6, 445, DFWB 7, 347f.

[7] DFWB 6, 4f.

[8] DFWB 7, 188.

[9] DFWB 7, 333ff.

bug(g)er, *humbug(g)en*, *humbugisch*[10] und Zusammensetzungen mit *Hiob-*[11] einiges gesagt.

All diese Nachweise sind an das Vorkommen geeigneter Belege geknüpft. Als aktuelles Beispiel für einige (nur historisch belegte) Ableitungen mag der Artikel „illuminieren" im derzeit in Arbeit befindlichen Band VIII des DFWB dienen: Unter „illuminieren" sind insgesamt 17 Sublemmata versammelt. Davon sind fünf nicht in den Korpora des IDS belegt: *Illuminant, illuminant, Illuminatist, Illuminatur* und *Illuminierer*, hinzu treten die selten belegten *illuminatisch, Illuminatismus, illuminatistisch, Illuminativ, illuminativ, illuminatorisch, Illuminismus, Illuminist* und *illuministisch*. Mit Hilfe von *Google Bücher* war es möglich, für alle diese Ausdrücke (z. T. lückenlose) Belegstrecken zusammenzustellen, einige der Ableitungen sind überhaupt erst mit Hilfe von *Google Bücher* belegbar.

4. Grenzen und Probleme

4.1 Ergebnisliste

Eine Suche mittels *Google Bücher* führt zur Ausgabe einer Ergebnisliste[12], die durch copyrightbedingte Einschränkungen in sich heterogen ist. Einige Bücher sind für alle Nutzer vollständig einseh- und benutzbar („vollständige Ansicht", diese Texte können zudem in Form von pdf-Dateien heruntergeladen werden), von anderen Texten werden nur aus-

---

[10] DFWB 7, 476ff.

[11] DFWB 7, 275ff.

[12] Im Folgenden wird versucht, den Ausdruck „Ergebnis" soweit möglich in derselben Weise zu verwenden, wie Google das tut, um den Bezug zu dem, was Google ein „Ergebnis" nennt, zu wahren. Damit sei aber keinesfalls zum Ausdruck gebracht, dass Google eine eindeutige, brauchbare oder auch nur auffindbare Definition von „Ergebnis" anbietet. Entsprechendes gilt im gleichen Fall für andere Bezeichnungen und Formulierungen, die Google Bücher verwendet, etwa „Wort", „Wortgruppe", „Buch" oder „Sortierung".

gewählte Seiten angezeigt („Vorschau"). Wenig brauchbar sind Ergebnisse, für die Google nur Minimalkontexte ausgibt, die in den meisten Fällen nicht einmal ganze Sätze beinhalten („Snippet-Ansicht") sowie Texte, zu denen lediglich einige (oft fehlerhafte oder unvollständige) bibliographische Daten vorhanden sind („Keine Vorschau verfügbar"). Ergebnisse in solchen Texten können durchaus brauchbar sein, müssen jedoch nötigenfalls per Hand nachgeschlagen werden.

Diese copyrightbedingten Einschränkungen können allerdings von vornherein in die Suche einbezogen werden, indem man bestimmte Textgruppen aus dem Gesamtbestand ausschließt. Dabei lässt sich beobachten, dass die ausgeschlossenen Textgruppen keine klar umgrenzten Teilmengen des Gesamtbestandes darstellen: Beschränkt man z. B. die Ergebnisliste mit Hilfe der Option „Nur vollständige Ansicht", werden oft weniger in vollständiger Ansicht vorhandene Bücher angezeigt, als tatsächlich im System verfügbar sind. Bisweilen gibt *Google Bücher* sogar an, es stünden überhaupt keine in vollständiger Ansicht vorhandenen Bücher zur Verfügung, obwohl tatsächlich und nachweislich mehrere solcher Bücher im System vorhanden sind: Lässt man sich nämlich sämtliche Ergebnisse ausgeben, finden sich darunter oft einige in vollständiger Ansicht verfügbare Bücher – oder doch mehr davon als zuvor. *Google Bücher* wählt also offenbar die Ergebnisse aus, die dem Nutzer präsentiert werden, ohne darauf aufmerksam zu machen und ohne die Auswahlkriterien offenzulegen.

4.2 Eingabe mehrerer Suchbegriffe

Bei jeder Suche mit *Google Bücher* muss bekanntlich berücksichtigt werden, dass die mittlerweile weithin übliche Trunkierung in diesem System nicht vorgesehen ist. Jede einzelne Flexionsform oder Schreibvariante muss im dafür vorgesehenen Feld „mit irgendeinem der Wörter" eigens eingegeben werden. Was Google nicht kommuniziert, ist, dass die Eingabe in diesem Feld auf 32 Formen beschränkt ist – was sich bei Suchen nach historischen Wortformen sehr schnell als zu limitiert erweisen kann. Man denke etwa an historische Schreibvarian-

ten wie *-ieren/-iren* bei Verben, die <k>/<c>-Varianz, die <I>/<J>-Varianz oder die <u>/<v>-Varianz.

Bei der Eingabe mehrerer Schreibvarianten bzw. Flexionsformen lässt sich gut beobachten, dass *Google Bücher* eine Auswahl trifft, wenn die Ergebnislisten zusammengestellt werden: So kann es geschehen, dass man bei einer Eingabe von einigen wenigen Wortformen Textstellen findet, die bei der Eingabe zusätzlicher Formen nicht mehr gefunden werden, auch wenn die ursprünglich eingegebenen Formen in diesen enthalten sind:

Eine Suche nach *illuminiren illuminieren* im Zeitraum zwischen 1740 und 1745 ergab am 18. November 2012 eine Liste von 167 Ergebnissen, eine Suche nach *illuminiren illuminieren illuminire illuminiere* im gleichen Zeitraum ergab 102 Ergebnisse. Unter diesen fehlten nicht nur 45 Ergebnisse, die die erste Suche zutage gefördert hatte, es kamen auch 12 Ergebnisse für *illuminiren illuminieren* hinzu, die bei der ersten Suche nicht gefunden worden waren.

Die Auswahl, die Google bei der Präsentation der Ergebnislisten trifft, führt also dazu, dass bereits gefundene Ergebnisse manchmal später nicht mehr auffindbar sind. In lexikographischen Arbeitsprozessen ist es aber unter Umständen nötig, einen Text, der als Beleg in Frage kommt, mehrfach aufzurufen. Selbstverständlich kann man den gefundenen Text in einem anderen Programm sichern – das verbessert jedoch nicht die Funktionalität von *Google Bücher*. Und was ist mit den Ergebnissen, die schon bei der ersten, entscheidenden Recherche vom System zwar gefunden, dem Benutzer aber nicht angezeigt wurden? Verlustig gegangene Belege lassen sich unter Umständen wiederfinden, z. B. durch das Wechseln des Rechners oder einen weiteren Versuch einige Tage später, in diesem Fall hat man aber schließlich Kenntnis von der Existenz des Gesuchten.

4.3 Wie arbeitet das System?

Die Situation wird dadurch verkompliziert, dass Google seit einiger Zeit eigenwillig zusätzliche Formen ausgibt, die gar nicht gefunden

werden sollten: Eine Suche nach *imprimirt bein* im Feld „mit allen Wörtern" vor 1800 am 14. 11. 2012 förderte auch Ergebnisse für *bien* und die folgenden Formen zutage: *imprimir T, imprimir. T, imprimir t/{, imprimir" t, imprimir t*[13] Selbstverständlich stand in den gefundenen Büchern nichts dergleichen, aber eben auch nicht die gesuchte Wortform *imprimirt*. Herausfiltern lassen sich solche Ergebnisse auch dann nicht, wenn man die von der OCR/ICR fehlinterpretierten Ausdrücke im Feld „ohne die Wörter" einträgt, vermutlich weil *Google Bücher* eine Form wie *imprimir. T* nicht als „Wort" ansieht. Das Beispiel zeigt jedoch, wozu eine derart großzügige Ergebnisausgabe führen kann: Bisweilen muss man sich durch lange Listen solcher unsinniger Ergebnisse kämpfen, um am Ende festzustellen, dass es nicht, wie in diesem Fall angegeben, 13, sondern überhaupt kein Ergebnis für die gesuchte Kombination gibt. Die Funktion „In den Ergebnissen suchen", die aus derart problematischen Ergebnislisten die wirklich interessanten Ergebnisse herausfiltern sollte, reproduziert lediglich dieselben Probleme.

Derzeit scheint *Google Bücher* zudem mit einer Art Semantisierung der Suche zu experimentieren: Eine Recherche am 27. September 2012 nach *Illuminant Jlluminant* im Feld „mit irgendeinem der Wörter" und *Bilder* im Feld „mit allen Wörtern" vor 1780 erbrachte fünf Ergebnisse (darunter Abb. 1–3).

---

[13] http://www.google.de/search?q=imprimiert+K%C3%B6rper&hl=de&safe=off&biw=1429&bih=906&sa=X&ei=Z7zhT_WD8LHsga1vZFx&ved=0CDQQpwUoBA&source=lnt&tbs=cdr%3A1%2Ccd_min%3A%2Ccd_max%3A31.12.1800&tbm=bks#q=imprimirt+bein&hl=de&safe=off&tbs=cdr:1,cd_max:31.12.1800&tbm=bks&psj=1&ei=tr_hT_CRO8XPsgbR-8Bw&start=10&sa=N&bav=on.2,or.r_gc.r_pw.r_qf.,cf.osb&fp=38bcbbf40bad90d2&biw=1429&bih=906

Hermanni Loemelii antuerpiensis sacrae theologiae ... Spongia, qua ... - Seite 94

books.google.de
German Loemelio - 1631 - 172 Seiten - Kostenloses Google eBook - Lesen
/4 prient , purgant , **illuminant** , ф perficíunt, & mue ft- tut em in purgatis lit ter is erudiunt , iuxta anliquatnßitu* ta i V tin ... quod dicitur, inquit, propter Curatos, veletiam **fotos** Titulares Epíícopos carête's plebe^quafidiceret, hos de Hierarchia ...

Abb. 1: Suche *Illuminant Illuminant* und *Bilder*

Hortus pastorum, sacrae doctrinae polymitus, authore R. D. Jacobo ... - Seite 549

books.google.de
Jacques Marchant, Alix - 1689 - 36 Seiten - Kostenloses Google eBook - Lesen
Et iîcut qui affiftunt , alij **illuminant** , alij purgant , ait) in mi- novos ac recentes femper parit frudtas, ita femper fion- ... con- tem hanc conftru&am eficnon folum pro Gentilibus ,' fluunt quicumque volnerint , сил íhbet **fotos** 'ordinii fed pro Filiis Ifrae'l. Mehr Ausgaben

Abb. 2: Suche *Illuminant Illuminant* und *Bilder*

Opera omnia: Band 2 - Seite 598

books.google.de
saint Bernard (de Clairvaux) - 1658 - Kostenloses Google eBook - Lesen
A.g.a Rccordatio peccatorum f*(e memoria diuinorum Cur **fotos** ineufet domeflicos , ab bxteri, *que Vniuerfados Jynagog* adnouam nuptam, ideft, beneficiorum eondienda. 3.14 a grauiter ... 5.67.a **illuminant** infert cordibut. A.11.1 71.a Dem ... Mehr Ausgaben

Abb. 3: Suche *Illuminant Illuminant* und *Bilder*

Hinter der Wortform *fotos* verbergen sich, soweit überhaupt lesbar, die lateinischen Formen *fons* und *solos*. Dies ist jedoch nicht der Punkt, auf den es in diesem Fall ankommt: Gesucht wurden Textstellen, an denen *Illuminant* oder *Jlluminant* in einem irgendwie gearteten Zusammenhang mit dem Wort *Bilder* auftreten. Gefunden wurden Ergebnisse, in denen *Illuminant* oder *Jlluminant* und das Wort *Fotos* zusammen auftreten. Steht dahinter der Versuch, eine Suche nach gleichbedeutenden Ausdrücken zu implementieren? Wie weit sind die diesbezüglichen Überlegungen und die technischen Ausarbeitungen vorangeschritten? Wird diese Suchmöglichkeit optional zu- bzw. abschaltbar sein? Und wird *Google Bücher* sich zu diesen Fragen äußern?

In Kenntnis des Systems steht zu befürchten, dass die Antwort auf die beiden letzten Fragen negativ ausfallen wird. Welche Auswirkun-

gen dies auf die Nutzbarkeit von *Google Bücher* haben könnte, lässt sich in diesem Stadium noch nicht absehen.

4.4 Die tatsächliche Anzahl der Ergebnisse

Seit einiger Zeit bietet *Google Bücher* die Möglichkeit, eine erzeugte Ergebnisliste chronologisch ordnen zu lassen. Die Funktion „nach Datum sortiert" ermöglicht es, auf diese Weise schneller und gezielter Zugriff auf Textstellen aus bestimmten Zeiträumen zu nehmen.[14] Auch hier kommt die Angewohnheit Googles, Ergebnisse auszuwählen, zum Tragen: Die Ergebniszahl ändert sich nämlich abhängig von der Sortierungsweise.

Eine Suche am 14. März 2012 nach *Imperialismen* im Feld „mit allen Wörtern" im Zeitraum zwischen 1850 und 1860 produzierte zwei Ergebnislisten, auf deren erster ein einziges Ergebnis angezeigt wurde (Abb. 4).

[14] Nebenbei bemerkt ist vollkommen unklar, was die zweite Sortierungsmöglichkeit, die bei Google „Nach Relevanz sortiert" heißt, leistet. Zwar hat es den Anschein, dass bei dieser Sortierungsweise deutsche Ergebnisse weiter „nach oben" sortiert werden, auf welche Weise das geschieht, ist allerdings unklar.

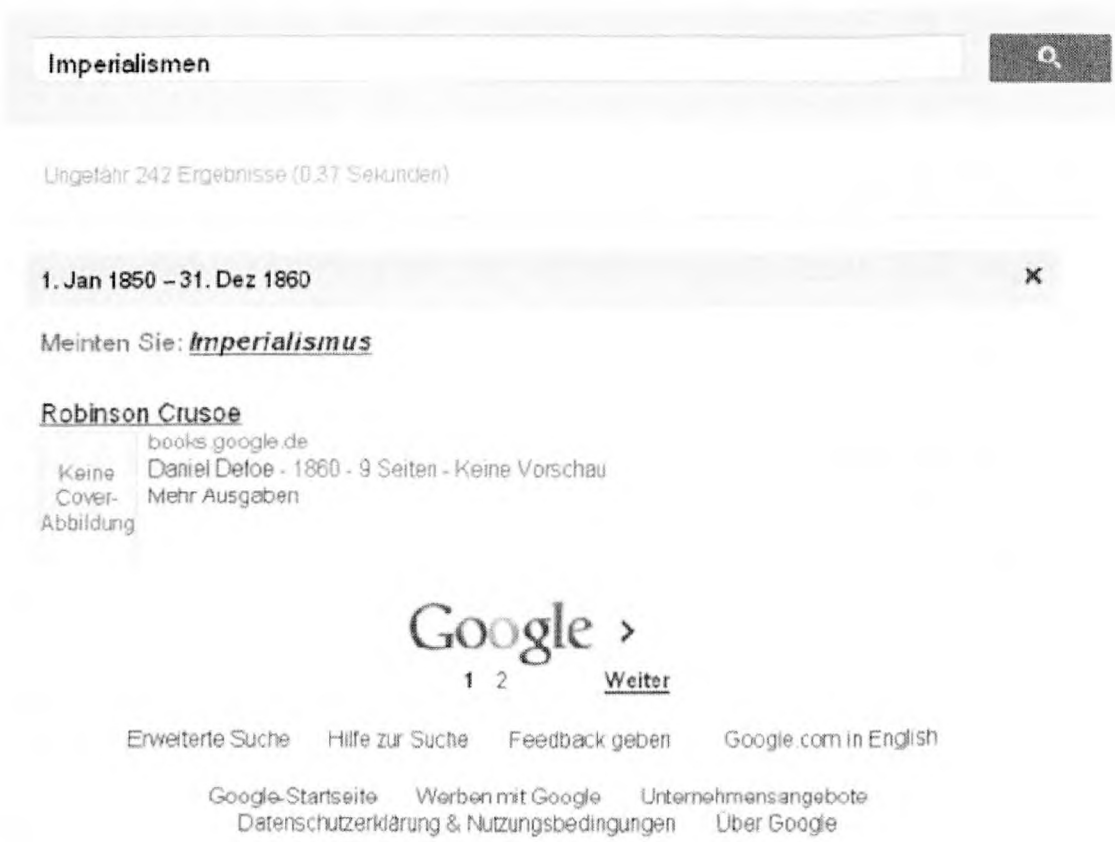

Abb. 4: Suche *Imperialismen zwischen 1850 und 1860*

Lässt man die Ergebnisliste darauf chronologisch sortieren, verschwindet der Titel. Die Gründe hierfür sind zwar unklar, das Phänomen ist deshalb aber nicht weniger interessant: Es lohnt sich nämlich demnach, die Ergebnisliste nach Datum bzw. Relevanz umsortieren zu lassen, auch wenn die jeweilige Sortierung gar nicht benötigt wird, da die Umsortierung eine neue Auswahl bedingt, durch die eventuell genau die Textstelle zutage gefördert wird, nach der man ansonsten lange vergeblich gesucht hätte.

Die chronologische Sortierung ist allerdings nicht besonders zuverlässig. Um das zu illustrieren, sei im Folgenden eine in BRÜCKNER (2012a, 17) dokumentierte Fallbeschreibung zitiert:

> Der Sucheintrag *Imperativismus Imperativismen* im Feld ‚mit irgendeinem der Wörter' mit der Vorgabe ‚vor 1980' erbrachte am 15.11.2010 eine Er-

gebnisliste, die 13 Einträge umfasste. Eine chronologische Sortierung der Liste mit Hilfe des Tools ‚nach Datum sortiert' generierte weitere fünf, also insgesamt 18 Ergebnisse, was wenig überraschend war, da dieser Effekt zu diesem Zeitpunkt bereits bekannt war. Als frühestes Ergebnis wurde in dieser chronologisch sortierten Liste eine Textstelle aus dem Jahr 1970 angegeben.

Nun wurde die gleiche Suche erneut vorgenommen, allerdings mit dem Unterschied, dass jetzt nur Textstellen vor 1900 ausgegeben werden sollten. Aufgrund der Erfahrungen aus der ersten Suche war allerdings zu erwarten, dass diese Suche kein einziges Ergebnis generieren würde. Tatsächlich gab Google Bücher aber fünf neue Ergebnisse aus, der früheste Beleg stammte diesmal aus dem Jahr 1876.

Welche Textstellen zwischen 1900 und 1970 hatte das System unterdrückt? Und warum?

Um dies zu überprüfen, wurde nun in einem nächsten Schritt erneut im Feld ‚mit irgendeinem der Wörter' nach *Imperativismus Imperativismen* gesucht, diesmal willkürlich zwischen 1890 und 1950. Diese Suche erbrachte (bereits chronologisch sortiert) 62 Ergebnisse, zwei Tage später (bereits chronologisch sortiert) 60 Ergebnisse und noch einmal eine ¾ Stunde später (bereits chronologisch sortiert) 59 Ergebnisse.[15]

In einem letzten Schritt wurde nun *Imperativismus Imperativismen* im gesamten Zeitraum zwischen 1891 und 1969 gesucht. Dies ergab zunächst 47, nach erfolgter chronologischer Sortierung 76 Ergebnisse. Das Resultat: Google unterschlug bei der ersten Suche (vor 1980) insgesamt mindestens 81 Ergebnisse, einschließlich des ‚Erstbelegs' (Suchen vor 1876 ergaben keine weiteren Ergebnisse – zumindest nicht an diesem Tag).

Wie viele Ergebnisse gibt es also wirklich? Eine Wiederholung der beschriebenen Suche nach *Imperialismen* zwischen 1850 und 1860 noch am gleichen Tag produzierte wieder die zwei Ergebnislisten, auf deren erster „Robinson Crusoe" angezeigt wird (Abb. 4). Interessant ist nun die Angabe „Ungefähr 242 Ergebnisse" oben. Es stellt sich nämlich die Frage nach den restlichen 241 Ergebnissen. Klickt man die zweite Ergebnisliste an, erhält man eine leere Seite mit dem Hinweis: „Es wurden keine Ergebnisse gefunden" (Abb. 5).

---

[15] Dass die Ergebniszahlen durch neu eingescannte Texte zunehmen, ist nachvollziehbar, dass sie aber abnehmen, muss dem Benutzer unverständlich bleiben.

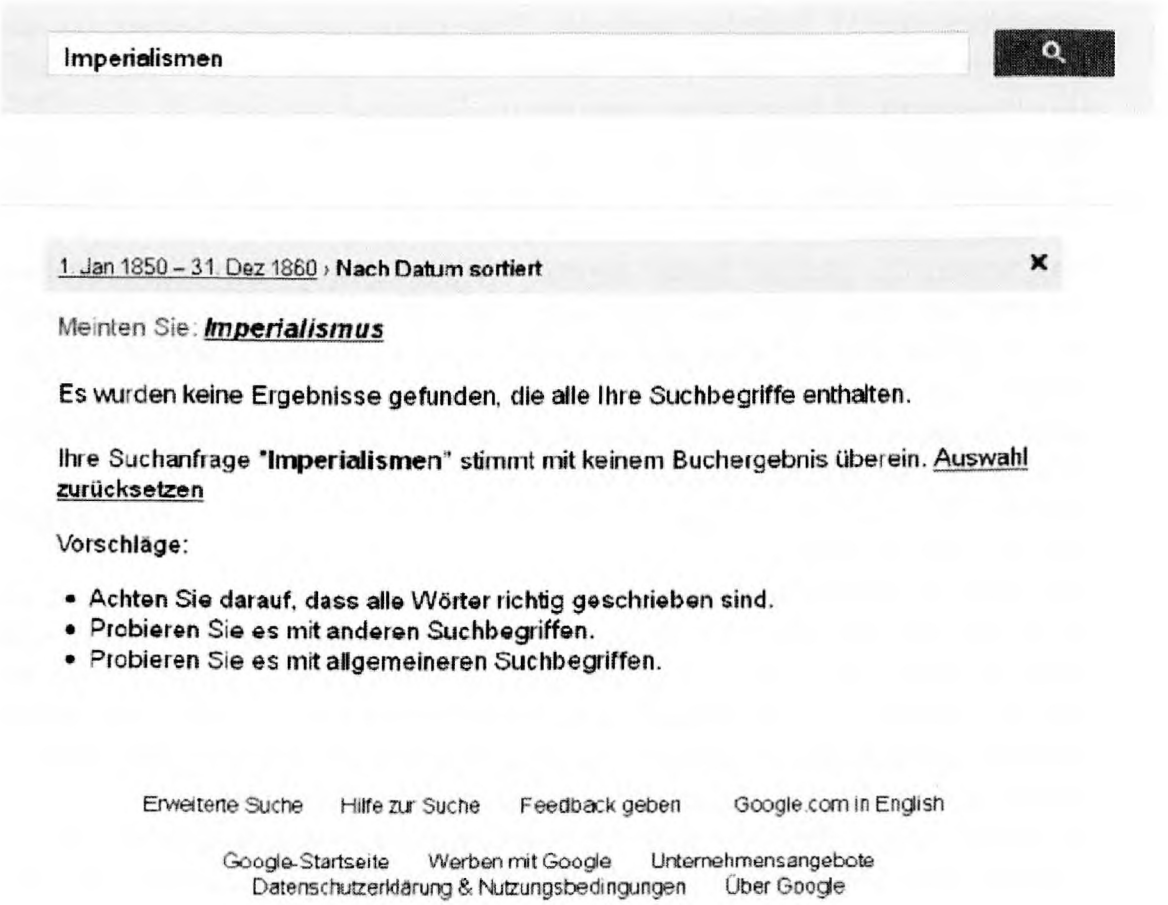

Abb. 5: Suche *Imperialismen zwischen 1850 und 1860* 2. Liste

Geht man einen Schritt zurück und lässt die angeblich 242 Ergebnisse umfassende Liste chronologisch sortieren, erhält man sodann wieder das Robinson-Crusoe-Ergebnis sowie die Angabe „Ungefähr 7 Ergebnisse“ (Abb. 6).

Abb. 6: Suche *Imperialismen zwischen 1850 und 1860 chronologisch*

Fälle wie dieser lehren, sich auf die Häufigkeitsangaben Googles nicht zu verlassen.

Gute Gründe, die dafür sprechen, mit den Ergebniszahlen schon auf Seiten Googles zurückhaltend zu sein, sind andernorts genannt (BRÜCKNER 2012a und 2012b), dazu gehören etwa das mehrfache Vorhandensein von Texten (etwa durch wiederholtes Einscannen des gleichen Texts, z. B. in verschiedenen Ausgaben, bei unterschiedlichen Versionen eines Texts oder Zitaten), Homographen (womöglich noch einzelsprachübergreifend) oder die bereits angesprochenen Verlesungen der OCR.

Exakte Ergebniszahlen zu liefern, ist vor diesem Hintergrund unmöglich. Dies erklärt Googles Zurückhaltung, die sich am greifbarsten noch in der Formulierung „ungefähr" ausdrückt, nicht jedoch das Zustandekommen der in unserem Beispiel mit 242 und 7 doch recht konkreten Zahlen. Wie kommen solche Angaben zustande, wenn nur ein einziges Ergebnis zu sehen ist? Warum ist dieses einzelne Ergebnis so

schwer zu zählen? Warum ändert sich die Zahl der Ergebnisse allein dadurch, dass man die Ergebnisliste wechselt? Warum ändert sich der Umfang der Liste von „ungefähr 242" auf „ungefähr 7" allein durch das Umsortieren – während das tatsächlich sichtbare Ergebnis, „Robinson Crusoe" sowohl identisch als auch alleine bleibt?

Die beschriebenen Probleme hängen offenbar mit zwei Grundproblemen zusammen: Das ist zum einen die Frage, was im Verständnis Googles ein Ergebnis ist, und zum anderen die Frage, auf welche Weise Google aus den im System vorhandenen Ergebnissen eine Auswahl trifft – und warum.

Die erste Frage ist von außen, ohne Kenntnis des Systems, nicht zu beantworten. Beobachtungen wie die, dass Google in den Ergebnislisten ein Buch, in dem das Suchwort ein einziges Mal vorkommt, ebenso als ein Ergebnis zu zählen scheint wie ein Buch, in dem das Suchwort mehrfach vorkommt[16], helfen diesbezüglich jedoch kaum weiter.

Was die Auswahl angeht, die Google trifft, so besteht das Hauptproblem darin, dass der Benutzer keinerlei Zugriff darauf nehmen kann. Weder kann er bestimmen, ob diese Auswahl überhaupt getroffen werden soll, noch kann er sie beeinflussen oder auch nur nachvollziehen.

Einen kleinen Einblick in das Funktionieren gewährt die folgende Beobachtung: Werden Ergebnismengen, die größer sind als 600, chronologisch geordnet, so fällt auf, dass aus ganzen Zeiträumen Ergebnisse fehlen; Ergebnisse, die auch dann nicht zu erlangen sind, wenn man sich durch sämtliche Ergebnislisten hindurchklickt. Der Zugriff auf die vermissten Ergebnisse ist nur durch eine erneute Suche möglich, bei der der Suchzeitraum entsprechend eingeschränkt wird, also so, dass die Suche nicht mehr als ca. 600 Ergebnisse generiert – was insbeson-

---

[16] Ähnliches gilt auch für in vollständiger Ansicht vorhandene Bücher, hier in durchaus sinnvoller Weise: Erscheint ein Wort z. B. in den Kopfzeilen mehrerer Seiten, etwa als lebender Kolumnentitel, so wird dieses Vorkommen zwar gelb unterlegt, aber nicht durch eine Markierung am rechten Rand als Ergebnis markiert.

dere bei hochfrequenten Suchwörtern äußerst mühsame Arbeiten nach sich ziehen kann.

Google Bücher scheint also die Ergebnislisten bei etwa 600 Einträgen zu „deckeln": Das bedeutet, dass Google maximal etwa 600 Ergebnisse auswählt, auch wenn erheblich mehr relevante Textstellen im System vorhanden und nachweislich verfügbar sind.

Da die Kriterien, nach denen *Google Bücher* diese Auswahl trifft, unklar bleiben und die Auswahl bei jeder erneuten Suche anders ausfällt (weshalb man häufig bereits gefundene Ergebnisse nicht wiederfindet), muss man mit Aussagen nicht nur über Häufigkeiten, sondern, gravierender, über das schiere Vorhandensein von Textstellen äußerst zurückhaltend sein. Allein auf Recherchen in *Google Bücher* kann man sich nicht stützen, schon gar nicht auf einmalige Suchen, die keine vergleichende Betrachtung ermöglichen. Zwar ist es möglich, einige Probleme mittels verschiedener Hilfskonstruktionen zu umgehen, da Google aber nicht offenlegt, wie die Buchsuche funktioniert, können diese nicht in einer Weise eingerichtet werden, die es gestatten würde, verlässliche Ergebnisse zu generieren.

5. Fazit

Die Verwendbarkeit der größten Sammlung digitalisierter Texte der Gegenwart wird durch eine Reihe von Faktoren erheblich eingeschränkt. Da sind zum einen technische Probleme, wie etwa die Qualität der OCR/ICR, die sich immerhin in den letzten Jahren zu verbessern scheint. Hinzu treten inhaltliche Schwächen wie etwa die (hier nur ganz am Rande gestreifte) Unzuverlässigkeit bibliographischer Daten.

Einer ganzen Reihe solcher Probleme könnte dadurch begegnet werden, dass Google dem Benutzer mehr Entscheidungsfreiheiten einräumt. Denn die Hauptursachen für die Unzuverlässigkeit des Systems liegen nicht in technischen Hemmnissen begründet, sondern in der Anlage eines Systems, das man jederzeit auch anders gestalten könnte. Gewährte man dem Nutzer Einblick in die Definitionen von Begriffen wie etwa „Wort", „Wortgruppe", „Buch", „Sortierung" oder

„Ergebnis" und überließe man ihm die Entscheidung darüber, ob, und wenn ja, wie das System in den verschiedenen Stadien einer Suche auswählt, wäre viel gewonnen – auch an Überprüfbarkeit von Daten minderer oder zweifelhafter Qualität, die auf ganz anderen Mängeln beruhen.

Die in BRÜCKNER (2009, 2012a und 2012b) vorgetragene Gesamteinschätzung ändert sich auch durch die neuen Erkenntnisse, die zum Teil auf Umstellungen im System von *Google Bücher* beruhen, nicht: Der immense Umfang der in *Google Bücher* zugänglich gemachten Textmenge macht es schwer, dieses Angebot zu ignorieren, *Google Bücher* kann aber allenfalls als ein Tool zum Einsatz kommen, das dem Lexikologen oder Lexikographen den Zugriff auf Texte und Textstellen erleichtert (oder ermöglicht), die er anderweitig nicht oder nur schwerlich hätte auffinden können. Als Quelle verlässlicher Daten, gar als Korpus kann es jedoch keinesfalls dienen. Es gilt auch weiterhin, dass es als ein gravierender Mangel des Systems anzusehen ist, dass bestimmte Probleme und offenkundige Fehler systembedingt sind und vom Nutzer in Form von Workarounds umgangen und korrigiert werden müssen. Es bleibt zu hoffen, dass Google sich irgendwann selbst an die Beseitigung dieser Mängel macht.

Literatur

BRÜCKNER, DOMINIK (2009): Die Google-Buchsuche als Hilfsmittel für die Lexikographie. In: Sprachreport 3/2009. Mannheim, 2009, 26–31.

BRÜCKNER, DOMINIK (2012a): Noch einmal: Die Google-Buchsuche. In: Sprachreport 2/2012. Mannheim, 16–20.

BRÜCKNER, DOMINIK (2012b): Google Bücher aus dem Blickwinkel des Lexikographen. In: Trefwoord, tijdschrift voor lexicografie. Jaargang 2012. Leeuwarden, Fryske Akademy.

COSMAS: http://www.ids-mannheim.de/cosmas2.

DFWB = Deutsches Fremdwörterbuch. Begonnen von HANS SCHULZ, fortgeführt von OTTO BASLER. 2. Auflage, völlig neu erarbeitet im Institut für Deutsche Sprache. Berlin/New York 1995ff.

GOOGLE-BÜCHER: http://books.google.de/advanced_book_search. Zur Diskussion über Google Bücher s. z. B.: JEANNENEY (2006). Zu einigen juristi-

schen Aspekten aus deutscher Sicht s. z. B. LEWANDOWSKI (2006 und 2008) sowie Weber (2010).

JEANNENEY, JEAN-NOËL (2006): Googles Herausforderung. Für eine europäische Bibliothek. Mit einem neuen Vorwort des Autors zur deutschen Ausgabe. Nachwort Klaus-Dieter Lehmann. Übersetzung: Sonja Finck, Nathalie Mälzer-Semlinger. Stiftung Preußischer Kulturbesitz Berlin. Berlin-Hamburg.

LEWANDOWSKI, DIRK (2006): Google Buchsuche. Bücher kostenlos zum Download. In: Password. 10/2006, 36.

LEWANDOWSKI, DIRK (2010): Wie verändert die Einigung mit Verlegern und Autoren die Buchwelt? In: Password 12/2008, 13.

WEBER, KLAUS (2010): Drei Jahre Freiheitsstrafe für alle Google-Mitarbeiter? Ein Beitrag zur Praxis des Urheberstrafrechts. In: Zeitschrift für Internationale Strafrechtsdogmatik, 220–226.